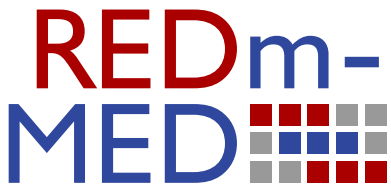

Review of Data Management Lifecycle Models

Alex Ball

redm1rep120110ab10.pdf

Issue date: 13th February 2012

Catalogue Entry

Title	Review of Data Management Lifecycle Models
Creator	Alex Ball (author)
Subject	research data management; research lifecycle; data lifecycle
Description	The requirements specified by the REDm-MED Project for data management plans will be assembled both from a 'bottom-up' elicitation process involving researchers and academics, and 'top-down' from funder requirements, institutional policy and known good practice. Data management lifecycle models provide a structure for considering the many operations that will need to be performed on a data record throughout its life; as such they can also act as a useful touchstone for ensuring the requirements specification has sufficient coverage of key data management operations.
Publisher	University of Bath
Date	10th January 2012 (creation)
Version	1.0
Type	Text
Format	Portable Document Format version 1.5
Resource Identifier	redm1rep120110ab10
Language	English
Rights	

Citation Guidelines

Alex Ball. (2012). *Review of Data Management Lifecycle Models* (version 1.0). REDm-MED Project Document redm1rep120110ab10. Bath, UK: University of Bath.

CONTENTS

1 INTRODUCTION

The approach being taken by the REDm-MED Project[1] towards the development of a Research Data Management Plan Requirements Specification is to extend the work completed by the ERIM Project[2] to a suit a wider range of users and needs. The *Engineering Research Data Management Plan Requirement Specification* [Bal+11] concentrated on planning for data re-use, data re-purposing and supporting data re-use. The remit of the REDm-MED version will be broader, taking in such topics as legal compliance, preservation, shareability and appraisal.

The earlier Requirements Specification will be extended in two ways. The first is through a requirements elicitation process involving researchers within the University of Bath's Department of Mechanical Engineering, using the DCC's CARDIO tool.[3] This represents the *practical* part of the approach. The second is through consideration of data management *theory*. This document summarizes a specific part of that theory for use in the process, namely models of the research data lifecycle; it is an updated version of Section 2.3 ('Data Lifecycles') of Ball [Bal10].

The importance of lifecycle models is that they provide a structure for considering the many operations that will need to be performed on a data record throughout its life. Many curatorial actions can be made considerably easier if they have been prepared for in advance – even at or before the point of record creation. For example, a repository can be more certain of the preservation actions it can perform if the rights and licensing status of the data has already been clarified, and researchers are more likely to be able to detail the methodologies and workflows they used if they record them at the time.

2 DCC CURATION LIFECYCLE MODEL

The DCC has produced a model for aligning curation tasks with the lifecycle stages of a digital object, intended as a planning tool for data creators, curators and users [Hig08]. A graphical representation of the model can be seen as Figure 1.

At the centre of the Model are the digital data, which are here identified with simple and complex digital objects or databases. The Model notes three levels of full-lifecycle actions:

- *Description* and (management of) *Representation Information*. The creation, collection, preservation and maintenance of sufficient metadata to enable the data to be used and re-used for as long as they have value to justify continued curation.
- *Preservation Planning*. Strategies, policies and procedures for all curation actions.
- *Community Watch and Participation*. The observation of what the OAIS Reference Model [CCS02] terms a Designated Community, that is, a predetermined group of stakeholders in the data, in order to track changes in their requirements for the data. Also, participation in the development of standards, tools and software relevant for the data.

1. REDm-MED Project Web page, URL: http://www.ukoln.ac.uk/projects/redm-med/
2. ERIM Project Web page, URL: http://www.bath.ac.uk/idmrc/erim/
3. CARDIO Website, URL: http://cardio.dcc.ac.uk/

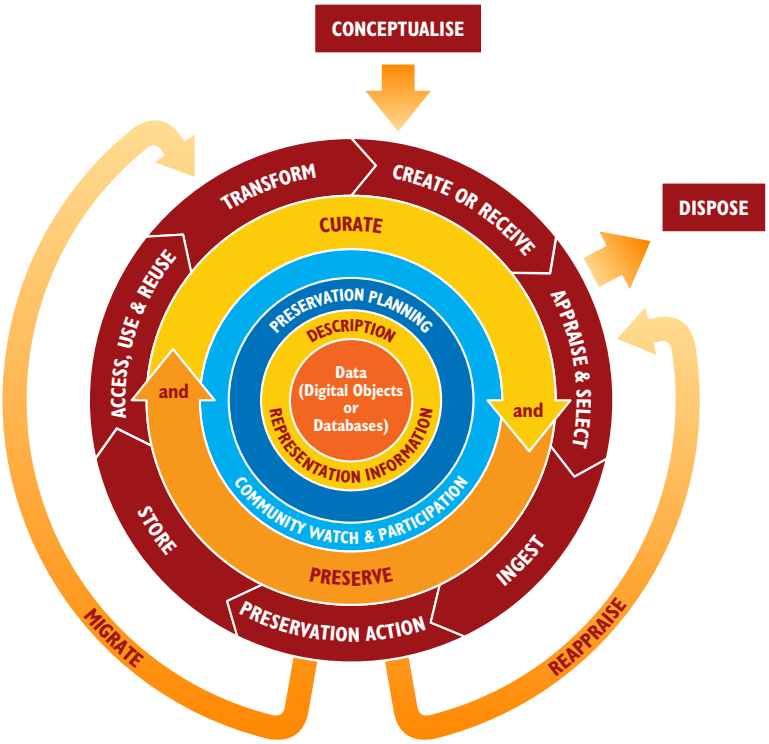

Figure 1: DCC Curation Lifecycle Model.

The fourth level, *Curate and Preserve*, properly describes most of the actions in the model, but is used here to represent the execution of the planned management and administrative actions supporting curation.

The sequential actions are not exclusively concerned with curation, but rather represent stages of the data lifecycle which ought to have a curation component. They begin with *Conceptualise*: the planning stages of the data generation and collection activities. Aspects such as the capture method will be informed by considerations other than curation – the scientific rigour of the method will be particularly important – but matters such as how the data will be stored, what budget to allocate for curation, and how collection of information important for curation may be automated or otherwise simplified, should be dealt with at this stage.

The curation lifecycle proper begins with the *Create or Receive* stage, where 'create' refers to original data generated and recorded by researchers, and 'receive' refers to pre-existing data collected from other sources. The curation activities at this stage centre on ensuring that all data is accompanied by sufficient administrative, descriptive, structural and technical metadata; ideally, pre-existing data should have these already, but as different researchers and repositories inevitably work to different standards they should be checked for consistency with local policies.

In the next stage, *Appraise and Select*, researchers or data specialists evaluate and select the data to keep for the long term according to documented guidance, policies or legal requirements. Some data may be sent for *Disposal*: this may involve transferring the

data to another custodian, although it could mean simple or secure destruction. Again, the nature of the disposal should be driven by documented guidance, policies or legal requirements. The remaining data are sent for *Ingest* by the normal custodian, be that an archive, repository, data centre or some other service. The Ingest stage immediately leads on to the *Preservation Action* stage, which involves an array of different activities: quality control, cataloguing, classifying, generating fixity data, registering semantic and structural metadata, and so on. Any data that fail quality control checks are returned to the originator for further appraisal. This should result either in improvements in the quality of the data (e.g. corrections to data transfer procedures, improved metadata, repackaging of data) and reselection, or disposal. Some data may need to be migrated to a different format, either to normalize it within the system or to reduce risks arising from hardware or obsolescence.

Once the data have completed the *Preservation Action* stage, they pass into *Storage*. This principally refers to the initial committal of the data to storage, but various long-term actions that ensure data remain secure may also be associated with this stage: maintaining the storage hardware, refreshing the media, making backup copies, checking for fixity, and so on.

Once the data have been safely stored, they enter a period of *Access, Use and Re-use*. Curation actions associated with this stage are focussed on keeping the data discoverable and accessible to designated users and re-users. This includes, for example, surfacing descriptive metadata through custom search interfaces or public APIs, and ensuring the preservation metadata held for the data continue to meet the requirements of the designated users and re-users.

Aside from ongoing preservation activities, the story of the archived data considered as an object stops at that point, but several events may cause progression to the *Transform* stage of the lifecycle. A key piece of software or hardware may approach obsolescence, therefore triggering an action to migrate the data to a new format, or a (re-)user may request a subset or other derivation of the data. The end result is a new set of data which starts the lifecycle again. Data created for a repository's internal purposes will, by its nature, pass through the early lifecycle stages rapidly, while data supplied to a (re-)user will progress as normal.

3 I2S2 IDEALIZED SCIENTIFIC RESEARCH ACTIVITY LIFECYCLE MODEL

The I2S2 Project produced a lifecycle model of scientific research from the researcher's point of view [I2S11]. As such it concentrates on the research, publication and administrative activities of the researcher, with only a high-level overview of archiving activity (which, for the most part, will be performed by a specialist). It was used by the I2S2 Project to identify gaps in the provision of tools to help automate, accelerate and integrate the research process.

A good place to begin reading the diagram of the model (see Figure 2) is in the top right-hand corner, with 'Research Concept and/or Experiment Design'. This is the point where researchers are considering a research question and how to answer it; any records from this stage are likely to be produced by hand. At the next stage, the researchers write a research proposal and a basic data management plan, most likely as electronic

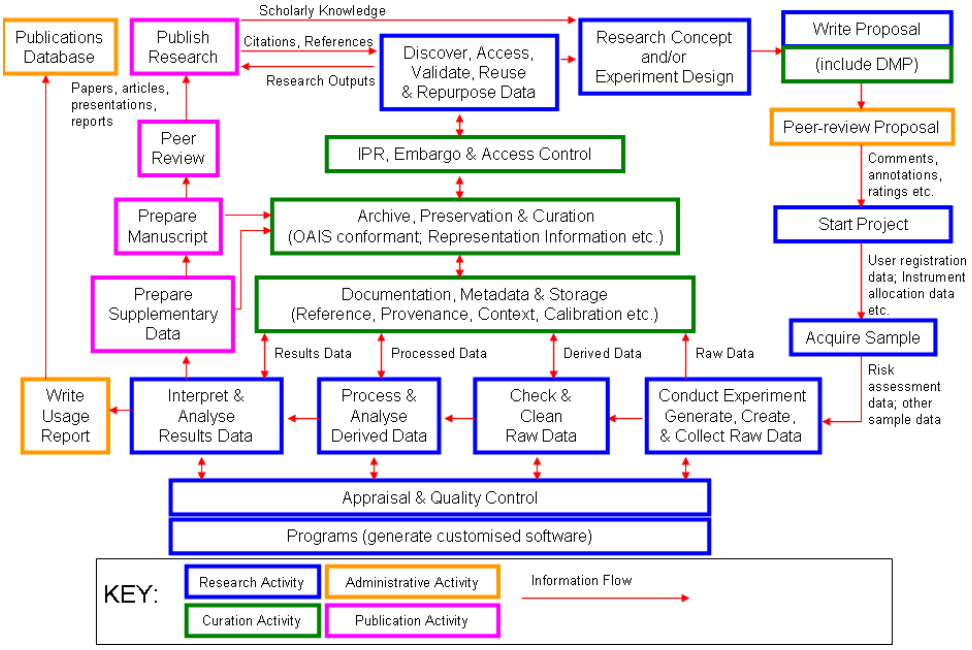

Figure 2: I2S2 Idealized Scientific Research Activity Lifecycle Model

documents. These documents are submitted to a funding body, which decides whether or not to fund the proposal based on peer reviews.

If the proposal is successful, the project starts. This stage may see the production of a fleshed-out project plan, a fuller data management plan, and the documentation of research instruments (whether to hand, acquired specially or, in such cases as questionnaires and interview templates, created by researchers). In some forms of research, additional submissions and approvals may be required before the research can proceed: for example, relating to health and safety or ethics. The next stage involves gaining access to (a sample of) the object of study. Once achieved, the researchers apply the research instruments to the object of study and generate data; they may also generate supplementary information and data providing context for the research.

Once the raw data are in hand, the researchers process them in various ways and analyse them (sometimes in an iterative fashion) in order to produce results data. Again, additional information and data about the workflow may be generated along the way. Finally, the researchers interpret the results data in order to answer the original research question. This concludes the research activity, but three other branches of activity continue:

- a publication activity (preparing a public version of the data, writing a journal paper, submitting both for publication);
- an administrative activity (writing final project reports for funders, writing internal usage reports, recording the publication activity);
- an archiving activity (most likely performed by a data archive or institutional data repository).

The life cycle completes when a future researcher consults the archived data and uses them as an inspiration for or input to a further research project.

4 DDI COMBINED LIFE CYCLE MODEL

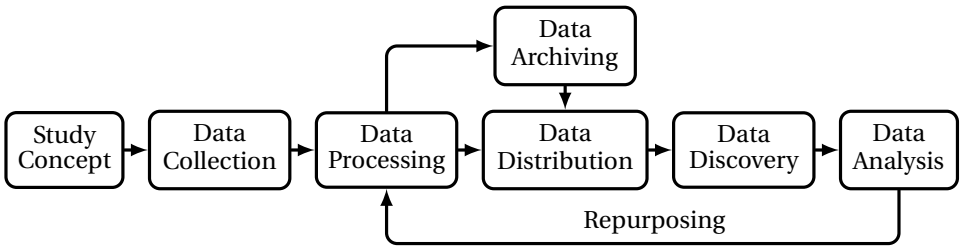

Figure 3: DDI version 3.0 Combined Life Cycle Model.

The Data Documentation Initiative (DDI) version 3.0 Conceptual Model [Str04] contains a Combined Life Cycle Model for research data, particularly social science data. The name signifies the fact that the model was formed by combining the initial draft of the model, constructed from a data application perspective, with elements from another model [GK02]. The model is linear for the most part, with one alternative path and one feedback loop; a graphical representation can be found as Figure 3. It has the following eight elements.

Study Concept. The earliest stage in the model is the point at which a survey is being designed. The model includes in this not only the choice of research question to answer and the methodology for collecting the requisite data, but also plans for the way in which the data will be processed, analysed and used to answer the question, and the form that answer will take. Researchers should also define at this stage the relationships that will exist between the data products of the research.

Data Collection. Examples given of collection methods and sources include surveys, censuses, voting or health records, commerce statistics or Web-based collections. Primary and secondary data sources should be clearly distinguished.

Data Processing. Once the input data has been assembled, it is processed and analysed to produce output data (e.g. a statistic or set thereof) that answers the research question. These output data may be recorded in a machine-readable form or human-orientated form such as a technical report.

Data Archiving. In order to ensure long-term access to the data, they should be passed to an archive rather than merely kept by researchers. The archive not only preserves the data (and metadata) but also adds value to them over time.

Data Distribution. The data are distributed to users either directly or via a library or data archive.

Data Discovery. The data may be publicised through books, journal publications, Web pages or other online services.

Data Analysis. The data may be used by others within the bounds of the original conceptualization; for example, picking out key statistics for a research report.

Repurposing. The data may also be used within a different conceptual framework; examples include sampling or restructuring the data, combining the data with other similar sets, or producing pedagogic materials.

Within the DDI standard, this model was used to group metadata requirements into five modules: study conception, data collection process, logical structure of encoded data, physical structure of encoded data, and archiving.

5 ANDS DATA SHARING VERBS

The Australian National Data Service (ANDS) has developed a set of Data Sharing Verbs as a model of the activities that need to be performed in order to make data available to a wider set of users [BT09]. They are used within the service as a structuring technique for operational planning and an advocacy tool for both data producers and data consumers. There are eight verbs in total.

Create. Refers to all the processes involved in creating a data object: generation, collection, aggregation, collation and transformation. During this stage, curation is supported principally by recording appropriate metadata at the earliest opportunity.

Store. The long term preservation of data within an archive, such as an institutional repository, institutional data store or national/international data centres.

Describe. The generation or acquisition of metadata supporting the storage, discovery, access and exploitation of data. These metadata should include descriptions of the people, organizations, and activities that lead to the creation of the data.

Identify. Assigning a persistent identifier to a set of data. ANDS provides a Handle-based identifier service, Identify My Data, for identifying data. Internationally, the DataCite Consortium provides services for registering DOIs for datasets [BF11].

Register. Making the existence of data known outside the context of its creation. There are various ways in which this may be achieved: providing a record for the data in a repository catalogue; providing this record in a machine readable format, perhaps using OAI-PMH or RSS, enabling it to be harvested and used in a repository cross-search service; citing the data from a published paper; publishing the data on the Web in RDF format.

Discover. Using a data discovery service to locate data of interest.

Access. Data may be accessed either through an automated system (i.e. using a Web link, perhaps passing through an authentication barrier and/or licensing agreement), or by application to a data custodian.

Exploit. Re-using data. The exploitation of data can only take place if good technical metadata (calibrations, classifications, metrics, etc.) are provided alongside contextual information about the way in which the data were created.

6 DATAONE DATA LIFECYCLE

The DataONE Federation uses a sequence of verbs, similar to the ANDS verbs (see Section 5), to describe the data lifecycle [MJ12]. It uses this lifecycle to place the tools from its toolkit into context.

1. *Plan.* Creating a data management plan to control how data are handled throughout the research project.
2. *Collect.* Acquiring data from sensors, laboratory instruments and measurement equipment in the field, and organizing them in spreadsheets or databases.
3. *Assure.* Agreeing standard formats, codes, units, etc. for the data and applying quality control procedures such as double entry, database input filters and outlier detection.
4. *Describe.* Providing sufficient metadata to ensure the data are independently understandable and reusable.
5. *Preserve.* Depositing data in a data centre or repository so they may be protected from bit-level degradation and format obsolescence.
6. *Discover.* Exposing discovery metadata through a searchable interface, and using such an interface to locate relevant data.
7. *Integrate.* Transforming several different datasets into a common representation (format, coding scheme, ontology), accounting for methodological and semantic differences while preserving a provenance trail.
8. *Analyze.* Applying statistical and analytical models to the data in order to extract meaningful answers to research questions.

The model actually combines several different workflows into a single expression. Typically, an investigation that only involves new data will proceed through steps 1 to 5 and then skip to step 8. A synthesis or meta-analysis might follow a cycle consisting of steps 6, 3, 7 and 8. In either case, there may be some overlap between the steps and fluidity to their order.

7 UK DATA ARCHIVE DATA LIFECYCLE

The UK Data Archive provides a data lifecycle model as an aid to researchers considering how data management relates to the lifecycle of a research project [UKD]. The model defines the following six stages.

1. Creating data
   - design research
   - plan data management (formats, storage etc)
   - plan consent for sharing
   - locate existing data
   - collect data (experiment, observe, measure, simulate)
   - capture and create metadata
2. Processing data
   - enter data, digitise, transcribe, translate
   - check, validate, clean data
   - anonymise data where necessary
   - describe data

    - manage and store data
3. Analysing data
    - interpret data
    - derive data
    - produce research outputs
    - author publications
    - prepare data for preservation
4. Preserving data
    - migrate data to best format
    - migrate data to suitable medium
    - back-up and store data
    - create metadata and documentation
    - archive data
5. Giving access to data
    - distribute data
    - share data
    - control access
    - establish copyright
    - promote data
6. Re-using data
    - follow-up research
    - new research
    - undertake research reviews
    - scrutinise findings
    - teach and learn

8 Research360 Institutional Research Lifecycle

The Research360 Project has developed a six-stage Institutional Research Lifecycle model [Jon11]. The model can be seen as a high-level summary of the I2S2 model (see Section 3), simplified to resemble the UKDA Lifecycle model (see Section 7). The main differences between the UKDA model and that of Research360 are that the latter has an additional planning stage at the start and no processing stage (see Figure 4).

9 Capability Maturity Model for Scientific Data Management

Crowston and Qin [CQ11] analysed data management literature to determine the key practices in the area, and then grouped them into four key process areas that form a short lifecycle. They did this to inform a new version of the Capability Maturity Model, as developed by the Software Engineering Institute at Carnegie Mellon University, dedicated to research data management.

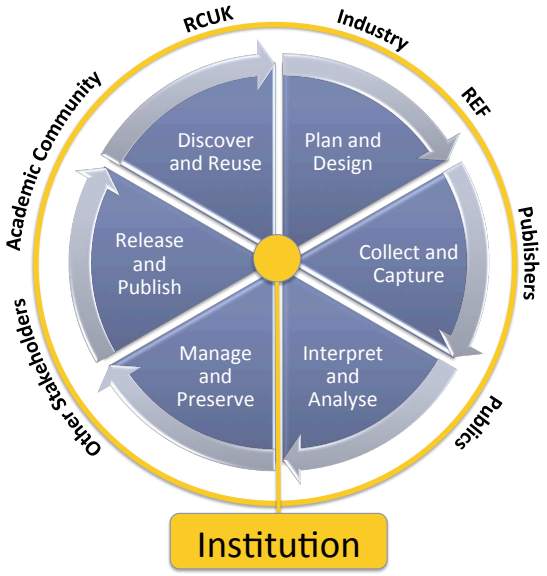

Figure 4: Research360 Institutional Research Lifecycle Concept

1. Data acquisition, processing and quality assurance
   Goal: Reliably capture and describe scientific data in a way that facilitates preservation and reuse
   a) Capture/acquire data
   b) Process and prepare data for storage, analysis and distribution
   c) Assure data quality (e.g. validate and audit data)
2. Data description and representation
   Goal: Create quality metadata for data discovery, preservation, and provenance functions.
   a) Develop and apply metadata specifications and schemas
   b) Contextualize, describe and document data
   c) Document data, software, sensors and mission
   d) Create descriptive and semantic metadata for datasets
   e) Design mechanisms to link datasets with publications
   f) Ensure interoperability with data and metadata standards
   g) Ensure compliance to standards
3. Data dissemination
   Goal: Design and implement interfaces for users to obtain and interact with data
   a) Identify and manage data products
   b) Encourage sharing
   c) Distribute data
   d) Provide access (e.g., by creating and piloting service models)

4. Repository services/preservation
   Goal: Preserve collected data for long-term use
   a) Store, backup and secure data (e.g., by backing up databases, preserving datasets and enforcing security of data systems)
   b) Manage schedules for archive generation, validation and delivery
   c) Curate data
   d) Perform data migration
   e) Build digital preservation network
   f) Validate data archives
   g) Package and deliver data archives

The maturity aspect of the model comes from considering how these practices are conducted for a given research project. If they are not performed at all, the project is at an unnamed zeroth level of maturity. If they are performed on an *ad hoc* basis, the project is at the *Initial* level of maturity. If they are performed according to policies and procedures set at the project level, the project is at the *Managed* level of maturity. If they are consistently applied across all projects in an institution, this means data management is performed at the *Defined* level of maturity. At the *Quantitatively Managed* level of maturity, data management processes are measured, evaluated and controlled, and once this control extends to constant improvement of effectiveness and efficiency, the final *Optimizing* level of maturity will have been reached.

In addition to the key practices, which represent the actual business of data management, Crowston and Qin also identified several practices that support the data management process. In the terms of the Capability Maturity Model, these are generic practices, grouped according to generic goals, that can be applied in any of the key process areas to increase its maturity. The generic practices that could move a project from the Initial stage of maturity to the Managed one include the following:

G2.1 The organization establishes policies for planning and performing the process
  G2.1.1 Develop data release policies
  G2.1.2 Develop sharing policies
  G2.1.3 Develop policies for data rights and rules for data use
  G2.1.4 Develop data curation policies
G2.2 A data management plan is established and maintained
  G2.2.1 Assess faculty data practices
  G2.2.2 Analyze data flows
  G2.2.3 Identify leading data management problems
  G2.2.4 Develop user requirements
  G2.2.5 Identify staffing needs
G2.3 Resources are provided
  G2.3.1 Develop business models for preservation

- G2.3.2 Appraise and evaluate enabling technologies (e.g., storage technology)
- G2.3.3 Develop SDM tools (e.g., tools to help researchers organize data, initial technology platforms or Web interface to the science repository)
- G2.3.4 Manage enabling technologies for access and conformance to standards

G2.4 Responsibility is assigned

- G2.4.1 Identify roles and responsibilities of personnel

G2.5 People are trained

- G2.5.1 Train researchers and data management personnel
- G2.5.2 Provide online guidance and workshops for data management

G2.6 Work products are controlled

- G2.6.1 Changes to data are controlled
- G2.6.2 Data provenance metadata is captured, including documentation of changes

G2.7 Stakeholders are identified

- G2.7.1 Develop collaboration and partnership with research communities and learned societies

G2.8 The process is monitored and controlled

- G2.8.1 Assess SDM effectiveness and impact

G2.9 Adherence to process standards is assessed and noncompliance addressed

- G2.9.1 Enforce policy

10 Conclusions

For the purposes of informing the REDm-MED Project's requirement specification for data management plans in the University of Bath's Department of Mechanical Engineering, the most useful lifecycle of those surveyed is probably that implied by the Capability Maturity Model for Scientific Data Management (see Section 9). This model explicitly aims to provide a relatively detailed and comprehensive list of the processes and practices necessary for a functioning and mature data management infrastructure, and assembles this list through reference to data science, data curation and data management literature. This being so, the model is well suited to act as a checklist of data management practices that merit attention in a data management plan, and which therefore should be borne in mind when laying out the requirements in the specification.

References

[Bal10] A Ball (2010-06-03). *Review of the State of the Art of the Digital Curation of Research Data*. ERIM Project Document erim1rep091103ab12. Version 1.2. University of Bath: Bath, UK. URL: http://opus.bath.ac.uk/19022.

[Bal+11] A Ball et al. (2011-01-17). *Engineering Research Data Management Plan Requirement Specification*. ERIM Project Document erim6rep100901ab11. Version 1.1. University of Bath: Bath, UK. URL: http://opus.bath.ac.uk/21280.

[BF11] J Brase & A Farquhar (2011). 'Access to Research Data'. *DLib Magazine* 17:1/2. DOI: doi:10.1045/january2011-brase.

[BT09] A Burton & A Treloar (2009). 'Designing for Discovery and Re-Use: the "ANDS Data Sharing Verbs" Approach to Service Decomposition'. *International Journal of Digital Curation* 4:3, 44–56. ISSN: 1746-8256. DOI: 10.2218/ijdc.v4i3.124.

[CCS02] Consultative Committee for Space Data Systems (2002). *Reference Model for an Open Archival Information System (OAIS)*. Blue Book CCSDS 650.0-B-1. Also published as ISO 14721:2003. URL: http://public.ccsds.org/publications/archive/650x0b1.pdf (2011-01-13).

[CQ11] K Crowston & J Qin (2011). 'A Capability Maturity Model for Scientific Data Management: Evidence from the Literature'. In: *Proceedings of the American Society for Information Science and Technology*. Vol. 48. DOI: 10.1002/meet.2011.14504801036.

[GK02] A Green & JP Kent (2002). 'The Metadata Life Cycle'. In: *MetaNet Work Package 1: Methodology and Tools*. Ed. by JP Kent. Chap. 2.2, 29–34. ISBN: 1-85764-017-9. URL: http://www.epros.ed.ac.uk/metanet/deliverables/D4/IST_1999_29093_D4.pdf (2010-03-11).

[Hig08] S Higgins (2008-07). 'The DCC Curation Lifecycle Model'. *International Journal of Digital Curation* 3:1, 134–140. ISSN: 1746-8256. DOI: 10.2218/ijdc.v3i1.48.

[I2S11] I2S2 Project (2011-04-07). *I2S2 Idealised Sceintific Research Activity Lifecycle Model*. URL: http://www.ukoln.ac.uk/projects/I2S2/documents/I2S2-ResearchActivityLifecycleModel-110407.pdf (2012-02-10).

[Jon11] K Jones (2011). *Research360: Managing data across the institutional research lifecycle*. Poster presented the 7th International Digital Curation Conference, Bristol, UK, 5–8 Dec. URL: http://www.dcc.ac.uk/webfm_send/589 (2012-02-13).

[MJ12] WK Michener & MB Jones (2012). 'Ecoinformatics: Supporting Ecology as a Data-Intensive Science'. *Trends in Ecology and Evolution* 27:2, 85–92. ISSN: 0169-5347. DOI: 10.1016/j.tree.2011.11.016.

[Str04] Structural Reform Group (2004-10-06). *DDI Version 3.0 Conceptual Model*. Data Documentation Initiative.

[UKD] *Research Data Lifecycle*. UK Data Archive. URL: http://www.data-archive.ac.uk/create-manage/life-cycle (2012-01-10).